

Multi-parameter One-Sided Monitoring Test

Guangyu Zhu*

Department of Statistics, University of British Columbia, Canada
Jiahua Chen

Research Institute of Big Data, University of Yunnan, China
Department of Statistics, University of British Columbia, Canada

March 16, 2017

Abstract

Multi-parameter one-sided hypothesis test problems arise naturally in many applications. We are particularly interested in effective tests for monitoring multiple quality indices in forestry products. Our search reveals that there are many effective statistical methods in the literature for normal data, and that they can easily be adapted for non-normal data. We find that the beautiful likelihood ratio test is unsatisfactory, because in order to control the size, it must cope with the least favorable distributions at the cost of power. In this paper, we find a novel way to slightly ease the size control, obtaining a much more powerful test. Simulation confirms that the new test retains good control of the type I error and is markedly more powerful than the likelihood ratio test as well as many competitors based on normal data. The new method performs well in the context of monitoring multiple quality indices.

Keywords: Bootstrap; Composite likelihood; Density ratio model; Empirical likelihood; Multiple sample; Random effect.

*The authors gratefully acknowledge funding from NSERC Grant RGPIN-2014-03743, a Collaborative Research and Development Grant from NSERC and FPInnovations, and the “a thousand talents” program through Yunnan University. We are also indebted to the Forest Products Stochastic Modelling Group centered at the University of British Columbia (UBC): members of this group from FPInnovations in Vancouver, Simon Fraser University, and UBC provided stimulating discussions of the long-term monitoring program to which this paper contributes.

1 Introduction

The research problem in this paper is motivated by an application. The reliability of a wood structure heavily depends on the mechanical strength of its component wood. It is important to closely monitor the dynamic wood strength distribution of solid lumber over time. This is done through data collected via a random sample from the target populations and the subsequent data analysis. A few weak components have potentially severe consequences for the structure, so the lower quantiles of the strength distribution have received the most attention. See the lumber-quality monitoring procedures specified in the American Society for Testing and Materials (ASTM) Standard D1990 (ASTM 1991). This is also evident from the recent report by Verrill et al. (2015), which examined the performance of various tests in the context of 5% quantiles.

Clearly, even if the strength distribution of the wood product meets the quality standard for the lower quantiles, the median or mean strengths could be significantly lower than the norm. The reliability of the structure could still be seriously compromised. This suggests the need to develop a monitoring test procedure for several quality indices simultaneously. We aim to draw the attention of practitioners to this need and to develop an effective and easy-to-use test procedure.

The application easily translates into a statistical question. We wish to statistically detect potential danger arising when the values of several user-selected parameters fall below well-established standards. In other words, we seek a test for multi-parameter one-sided null and alternative hypotheses. More abstractly, suppose we have a sample from distribution F , and θ is a vector-valued parameter or functional of F . We wish to test the hypothesis

$$H_0 : \theta \geq \theta^* \text{ against the alternative } H_a : \theta \not\geq \theta^* \quad (1)$$

for a specific known vector $\boldsymbol{\theta}^*$, where the inequality is interpreted to be component-wise. Because of the invariance property, without loss of generality, we may take $\boldsymbol{\theta}^* = \mathbf{0}$; this will be assumed hereafter unless otherwise indicated. The dimension of $\boldsymbol{\theta}$ will be denoted as p . Clearly, many existing tests can easily be adapted to this problem. However, we suggest that none of them seem to exactly fit, and additional research is needed.

Under the normal model, the likelihood ratio test (LRT) provides standard solutions to the current pair of opposing hypotheses H_0 and H_a and similarly formulated pairs of opposing hypotheses. Statisticians must determine the appropriate rejection region to ensure that the LRT has the size specified by the user. Along this line, Robertson & Robertson (1988) worked out the solution to the LRT problem for the case where $\boldsymbol{\Sigma}$ is known to be \mathbf{I} . Perlman (1969) solved the LRT problem where $\boldsymbol{\Sigma}$ is unknown.

By the standard definition in mathematical statistics, the size of a test is the supremum of its type I error. When the null hypothesis is composite, i.e., it contains many distributions, the size of the test is the type I error in the worst scenario, or at the least favorable null distribution. Controlling the size of the test can therefore lead to a pessimistic procedure: the type I error under the likely true data-generating distribution is far below the size of the test that leads to compromised power. This is particularly true for the LRT for multi-parameter one-sided hypotheses. Perlman & Wu (2003) and Perlman & Wu (2006) examined the rejection region of the LRT in many situations and developed more powerful tests accordingly. Such research is often motivated by medical studies, where the aim is often to assess whether a therapy has a beneficial effect on multiple outcomes simultaneously relative to a control. The specifics of these one-sided hypotheses vary depending on the medical problem. For instance, O'Brien (1984) and Tang et al. (1989) proposed and extended a generalized least-squares test that is most powerful when the true population mean is near a specific line in the alternative space. In clinical studies with

multiple outcomes, researchers may wish to confirm that a new treatment is superior in at least one of the outcomes and equivalent on the rest of the outcomes, in comparison with the control. Tamhane & Logan (2004) targeted this problem with a test derived from the union–intersection test of (Roy 1953) and the intersection–union test of (Berger 1982). We refer to Wassmer et al. (1999) for a more detailed review of this area and Lachin (2014) for recent advances.

The hypothesis of interest in this paper, (1), is similar to but different from those considered in the above papers. We investigate the direct application of the standard LRT to (1) and discover that a specific version of the LRT leads to a much improved procedure that is particularly useful for our application. We find a novel way to mildly relax the size control to obtain a much more powerful test. Simulation confirms that the new test retains tight control of the type I error and is markedly more powerful than the LRT as well as many of its competitors based on normal data. The new method performs well in the context of monitoring multiple quality indices.

The paper is organized as follows. In Section 2, we revisit some basics of the LRT, introduce the new test, and review existing methods for normal data and one-sided multi-parameter hypotheses. In Section 3, we give a brief background on the monitoring test for forestry products and the application of the proposed method. In Section 4, we present simulation results. We conclude in Section 5.

2 Proposed and related methods

The new approach was developed as a result of our observation of the LRT under the normal model. For this reason, we first quickly revisit the standard likelihood approach and then introduce our approach.

2.1 LRT statistic

Suppose we have an independent and identically distributed (iid) sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ from a p -dimensional multi-normal distribution $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We first consider the test problem for

$$H_0 : \boldsymbol{\mu} \leq 0 \text{ against } H_1 : \boldsymbol{\mu} \not\leq 0. \quad (2)$$

Let \mathbf{X} denote the sample mean $\bar{\mathbf{Y}}$ and

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T,$$

a slightly altered sample variance. It is well known that \mathbf{X} and \mathbf{S} together are complete and sufficient for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ under the normal model. Hence, we may develop a likelihood-based method as if they are the only observations.

After some simple algebra, the log-likelihood function is found to be

$$\ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} \log \det(\boldsymbol{\Sigma}) - \frac{n}{2} \text{tr}\{\boldsymbol{\Sigma}^{-1}[\mathbf{S} + (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]\}.$$

To develop an LRT, we search for the maximum point of $\ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ under the null hypothesis and under the full model. The solution under the full model is well known, with the unconstrained maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given by

$$\hat{\boldsymbol{\mu}} = \mathbf{X}; \quad \hat{\boldsymbol{\Sigma}} = \mathbf{S}.$$

This implies

$$\sup \ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -(n/2)\{\log \det(\mathbf{S}) + p\}.$$

The solution under the null model is algebraically simple but slightly more abstract. For each fixed $\boldsymbol{\mu}$, we find

$$\arg \max_{\boldsymbol{\Sigma}} \ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \mathbf{S} + (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T.$$

This leads to the profile log-likelihood function of $\boldsymbol{\mu}$:

$$\begin{aligned}\ell_n(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_\mu) &= -\frac{n}{2}\{\log \det(\mathbf{S} + (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T) + p\} \\ &= -\frac{n}{2}\{\log \det(\mathbf{S}) + \log[1 + (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{X} - \boldsymbol{\mu})] + p\}.\end{aligned}\quad (3)$$

The second equality is obtained by a linear algebra result $\det(\mathbf{I} + \mathbf{u}\mathbf{v}^T) = 1 + \mathbf{u}^T \mathbf{v}$ for any vector \mathbf{u} and \mathbf{v} , and by

$$\mathbf{S} + (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T = \mathbf{S}^{1/2}\{\mathbf{I} + [\mathbf{S}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})][\mathbf{S}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})]^T\}\mathbf{S}^{1/2}.$$

Clearly, the profile likelihood is maximized if and only if $(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ is minimized with respect to $\boldsymbol{\mu}$ in the space of the null hypothesis. Let the solution to the minimization problem be $\hat{\boldsymbol{\mu}}_0$. Geometrically, it is the projection of \mathbf{X} onto the null space in terms of the Mahalanobis distance defined through the covariance matrix \mathbf{S} . Subsequently, we find the generic expression of the LRT statistic:

$$R_n = 2\{\sup \ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \sup_{H_0} \ell_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} = n \log\{1 + (\mathbf{X} - \hat{\boldsymbol{\mu}}_0)^T \mathbf{S}^{-1}(\mathbf{X} - \hat{\boldsymbol{\mu}}_0)\}.$$

Note that R_n is monotonic in

$$T_n = n(\mathbf{X} - \hat{\boldsymbol{\mu}}_0)^T \mathbf{S}^{-1}(\mathbf{X} - \hat{\boldsymbol{\mu}}_0). \quad (4)$$

Thus, the rejection region of the LRT statistic has the generic form

$$C = \{(\mathbf{Y}_1, \dots, \mathbf{Y}_n) : T_n > c\} \quad (5)$$

for some c , which is called the critical value of the test.

By classical theory in mathematical statistics, if the size of the test is set to α , then the critical value c will be chosen so that

$$\sup_{\boldsymbol{\mu} \in H_0, \boldsymbol{\Sigma} > 0} \Pr\{T_n > c; \boldsymbol{\mu}, \boldsymbol{\Sigma}\} = \alpha \quad (6)$$

where we use $\Pr(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to indicate that the calculation is under the $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. According to Perlman (1969), the supremum is attained asymptotically when $\boldsymbol{\mu} \rightarrow 0$ and $\boldsymbol{\Sigma}$ approaches some singular matrix. Specifically, he proved that for H_0 defined by (1),

$$\begin{aligned} \sup_{\boldsymbol{\mu} \in H_0, \boldsymbol{\Sigma} > 0} \Pr\{T_n > c; \boldsymbol{\mu}, \boldsymbol{\Sigma}\} &= \frac{1}{2} \Pr[\mathbf{F}_{p-1, n-p+1} \geq (\frac{1}{p-1} - \frac{1}{n})c] \\ &+ \frac{1}{2} \Pr[\mathbf{F}_{p, n-p} \geq (\frac{1}{p} - \frac{1}{n})c] \end{aligned} \quad (7)$$

where $\mathbf{F}_{p,n}$ denotes an F-distributed random variable with p and n degrees of freedom. In other words, an LRT of size α will choose c such that

$$\Pr[\mathbf{F}_{p-1, n-p+1} \geq (\frac{1}{p-1} - \frac{1}{n})c] + \Pr[\mathbf{F}_{p, n-p} \geq (\frac{1}{p} - \frac{1}{n})c] = 2\alpha. \quad (8)$$

2.2 Proposed test

The choice of c in the LRT in (8) ensures that the type I error is at most α at any $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in H_0$. When the dimension of the data $p = 2$, the type I error is maximized when $\boldsymbol{\mu} = \mathbf{0}$ and $\rho \rightarrow -1$ where ρ is the correlation coefficient. If the observations are from a distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\rho = 0$, the type I error is far lower than α . In many applications, the user may be confident that $\rho \geq 0$. If so, this choice is far too conservative. The size of the test over the region of interest is much lower than the designated α . As a consequence, the power of the test is also much lower.

This consideration begs a question on the type I error of the test at $\boldsymbol{\mu} = 0$ and a given $\boldsymbol{\Sigma}$. Interestingly, an answer is readily available from Nüesch (1966). To state this result, we first introduce some notation. When \mathbf{X} is $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we use the simplified notation

$$\Pr\{\boldsymbol{\Sigma}\} = \Pr\{\mathbf{X} > 0; \boldsymbol{\mu} = 0, \boldsymbol{\Sigma}\}.$$

Let \mathcal{S} be the collection of all nonempty subsets of $\{1, 2, \dots, p\}$. We use $\mathbf{X}[i]$ for the i th entry of vector \mathbf{X} . For any $s \in \mathcal{S}$, we use $\mathbf{X}[s]$ for the subvector of \mathbf{X} consisting of

components of $\mathbf{X}[i]$ such that $i \in s$. Let s' be the complement of s . With these, we use Σ_s for the covariance matrix of $\mathbf{X}[s]$ and $\Sigma_{s'|s}$ for the covariance matrix of $\mathbf{X}[s']$ conditional on $\mathbf{X}[s] = 0$. We use the convention that when s' is empty $\Pr\{\Sigma_{s'|s}\} = 1$. We use $|s|$ for the size of s . In the following theorem, T_n is the LRT statistic defined earlier.

Theorem 1. *In the current setting, for any $c > 0$,*

$$\Pr(T_n > c) = \sum_{s \in \mathcal{S}} \Pr\{\mathbf{F}_{|s|, n-|s|} > \left(\frac{1}{|s|} - \frac{1}{n}\right)c\} \Pr\{\Sigma_s^{-1}\} \Pr\{\Sigma_{s'|s}\}.$$

In other words, the distribution of T_n is a finite mixture of F -distributions. The proof of this theorem is technically involved; we refer to Nüesch (1966) for the details.

The probabilities in the above theorem have generic analytical expressions that can be found in Kendall (1941). We are particularly interested in the case $p = 2$. When $p = 2$, without loss of generality, we assume that X has marginal variances 1 and denote the correlation coefficient as ρ . For s such that $|s| = 1$, it is easy to see that

$$\Pr\{\Sigma_s^{-1}\} = \Pr\{\Sigma_{s'|s}\} = \frac{1}{2}.$$

When $|s| = 2$, the correlation coefficient specified by Σ^{-1} is $-\rho$. Let Z_1, Z_2 be two independent $N(0, 1)$ random variables. Then, $X_1 = Z_1$ and $X_2 = \sin(\gamma)Z_2 - \cos(\gamma)Z_1$ have correlation $-\rho$ when $\gamma = \arccos(\rho)$ in the range of 0 and π . Hence,

$$\Pr\{\Sigma_s^{-1}\} = \Pr(Z_1 > 0; \sin(\gamma)Z_2 - \cos(\gamma)Z_1 > 0) = \frac{\gamma}{2\pi}.$$

In other words, we have

$$\Pr(T_n > c) = \frac{1}{2} \Pr\{\mathbf{F}_{1, n-1} \geq \left(1 - \frac{1}{n}\right)c\} + \frac{\arccos(\rho)}{2\pi} \Pr\{\mathbf{F}_{2, n-2} \geq \left(\frac{1}{2} - \frac{1}{n}\right)c\}. \quad (9)$$

Consequently, if the value of ρ is known and the observed value of T_n is t_{obs} , we would have evaluated the p value of the test to be

$$\frac{1}{2} \Pr\{\mathbf{F}_{1, n-1} \geq \left(1 - \frac{1}{n}\right)t_{obs}\} + \frac{\arccos(\rho)}{2\pi} \Pr\{\mathbf{F}_{2, n-2} \geq \left(\frac{1}{2} - \frac{1}{n}\right)t_{obs}\}.$$

This would lead to a much more powerful test than the classical LRT. For instance, we would reject H_0 when $t_{obs} = 4.59$ when ρ is known to be 0, while the LRT does not reject in this case. See Table 1 for the critical values. The LRT uses the critical value at $\rho = -1$, corresponding to the least favorable distribution.

Table 1: Critical values of the LRT test when ρ is known and $n = 50, p = 2$.

ρ	-1.0	-0.9	-0.5	0	0.5	0.9
c	5.64	5.37	4.98	4.58	4.12	3.47

Motivated by the above discussion and calculations, we propose a new test for $p = 2$. First, we obtain the value of T_n and the sample correlation coefficient $\hat{\rho}$. With the observed value t_{obs} , we compute

$$\hat{p} = \frac{1}{2} \Pr\{\mathbf{F}_{1,n-1} \geq (1 - \frac{1}{n})t_{obs}\} + \frac{\arccos(\hat{\rho})}{2\pi} \Pr\{\mathbf{F}_{2,n-2} \geq (\frac{1}{2} - \frac{1}{n})t_{obs}\}. \quad (10)$$

The test rejects H_0 when $\hat{p} < \alpha$, where α is the designated size of the test.

Our idea is not limited to $p = 2$. The analytical form of \hat{p} (the p-value of the test) is more complex in the general case but can be calculated according to Theorem 1. We do not present the details here since the interested user can work them out with some algebraic effort. We call the new test the mLR test.

The type I error of the mLR test may in theory exceed α at some specific ρ values very close to -1 . Our simulation experiments show that the degree of inflation is negligible.

2.3 Application to non-normal data

In applications, the data are often collected from non-normal populations. Nevertheless, it is generally possible to obtain a good estimate of the vector parameter $\boldsymbol{\theta}$ of dimension p

and its covariance matrix. We consider the situation where

$$\sqrt{n}\mathbf{S}_n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow \text{MVN}(0, \mathbf{I})$$

in distribution when some index, likely the sample size n , goes to infinity.

Suppose it is of interest to test the hypothesis in the form of (1) and, without loss of generality, $\boldsymbol{\theta}^* = \mathbf{0}$. The proposed modified LRT can be applied to this problem by setting $\mathbf{X} = \hat{\boldsymbol{\theta}}$ and $\mathbf{S} = \mathbf{S}_n$. The computation of T_n and \hat{p} can then be carried out in the same way. We reject the null hypothesis when $\hat{p} < \alpha$. When the sample size n is large, one may use χ_p^2 to replace $\mathbf{F}_{p,n}$ and so on to give an approximate \hat{p} .

2.4 Other methods

As pointed out earlier, there exist many methods to handle the hypothesis test problem under a multivariate normal model. It is helpful to see how the proposed method differs. For brevity, we give a quick introduction to just two methods. We still assume that an iid sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ from $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given and will continue to use some of the notation introduced earlier.

Union–Intersection Test In the union–intersection test (UIT), we start by defining sub-null hypotheses $H_{0j} = \{\boldsymbol{\mu} : \mu_j \leq 0\}$ for $j = 1, 2, \dots, p$. Clearly, $H_0 = \bigcap_{j=1}^p H_{0j}$. This means that if any H_{0j} is false, then H_0 is also false. Thus, one may test the validity of H_{0j} for each j . We reject H_0 if any H_{0j} is rejected.

When $\boldsymbol{\Sigma}$ is known to be \mathbf{I} , we may reject H_{0j} when the component sample mean of the j th component $\bar{\mathbf{Y}}_j > c$ for some critical value $c > 0$. We reject H_0 when

$$\max\{\bar{\mathbf{Y}}_j : j = 1, \dots, p\} > c.$$

Note that under the null hypothesis

$$\begin{aligned}\Pr(\max\{\bar{\mathbf{Y}}_j, j = 1, \dots, p\} > c) &= 1 - \Pr(\max\{\bar{\mathbf{Y}}_j, j = 1, \dots, p\} < c) \\ &= 1 - \prod_{j=1}^p \Pr(\bar{\mathbf{Y}}_j < c).\end{aligned}\tag{11}$$

Hence, we may choose $c = z_{(1-\alpha)^{1/p}}/\sqrt{n}$ to obtain a size α test, where $z_{(1-\alpha)^{1/p}}$ is the lower $(1 - \alpha)^{1/p}$ quantile of the standard normal distribution.

When Σ is unknown, we may conduct a one-sided t -test of size α/p for H_{0j} for $j = 1, 2, \dots, p$. We reject H_0 when any H_{0j} is rejected. By the Bonferroni inequality we see that the size of this test below α . It is well known that a test formed by Bonferroni correction tends to be very conservative.

PW test. Perlman & Wu (2003) were among the first to take note of the conservative nature of both UIT and LRT. In particular, they suggested that the boundary of H_0 can be decomposed into subspaces of varying dimensions. For instance, when $p = 2$, the boundary of $\{\boldsymbol{\mu} \leq 0\}$ is decomposed into

$$B_1 = \{\mu_1 = 0, \mu_2 = 0\}, B_2 = \{\mu_1 < 0, \mu_2 = 0\}, B_3 = \{\mu_1 = 0, \mu_2 < 0\}.$$

The dimension of B_1 is 0 and that of B_2 and B_3 is 1. If the sample mean $\mathbf{X} \in H_0$, then $T_n = 0$. Otherwise, the maximum of the distances from \mathbf{X} to B_1 , B_2 , or B_3 is taken as T_n . The information on the source of the maximum is then discarded, and the size of T_n is measured against the least favorable distribution, which corresponds to $\boldsymbol{\mu} \in B_1$ and $\rho = -1$.

Perlman & Wu fix the conservative nature of the LRT by having different critical values

depending on the location of \mathbf{X} with respect to B_1 , B_2 , or B_3 . Let

$$\begin{aligned} M_1 &= \{n\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X} > c_{2,\alpha}\} \\ M_2 &= \left\{\frac{\mathbf{X}_1}{\sqrt{\mathbf{S}_{11}/(n-1)}} > t_{n-1,\alpha}\right\} \\ M_3 &= \left\{\frac{\mathbf{X}_2}{\sqrt{\mathbf{S}_{22}/(n-1)}} > t_{n-1,\alpha}\right\}, \end{aligned}$$

where $c_{2,\alpha}$ is the critical value of the LRT test of size α , according to (7), and \mathbf{S}_{11} and \mathbf{S}_{22} are entries of matrix \mathbf{S} . The PW test (Perlman & Wu 2006) rejects H_0 when $\mathbf{X} \in M_1 \cap (M_2 \cup M_3)$. That is, H_0 is rejected when B_1 is rejected and one of B_2 and B_3 is also rejected.

We can verify that the rejection region of the PW test covers the rejection region of the LRT; see Figure 1. At the least favorable distribution where $\rho = -1$, its type I error will exceed α , as is the case for our method. When $\rho = -0.9$ the type I error of the PW test is 5.46% based on our simulations.

3 Application to monitoring test

The proposed modified LRT is developed with an application in mind. As discussed by Verrill et al. (2015), forestry is concerned with monitoring the lower quantiles of the mechanical strength distribution. Many researchers focus on the 5th quantile. In this paper, we simultaneously monitor several quality parameters of the mechanical strength distribution. In this section we demonstrate the usefulness of the modified LRT.

The modified LRT may be used in many ways and many applications. We, however, focus on the specific setting and inference methods developed in Chen et al. (2016). We refer to this paper for more detailed background information but provide some necessary description of the data and inference methods here.

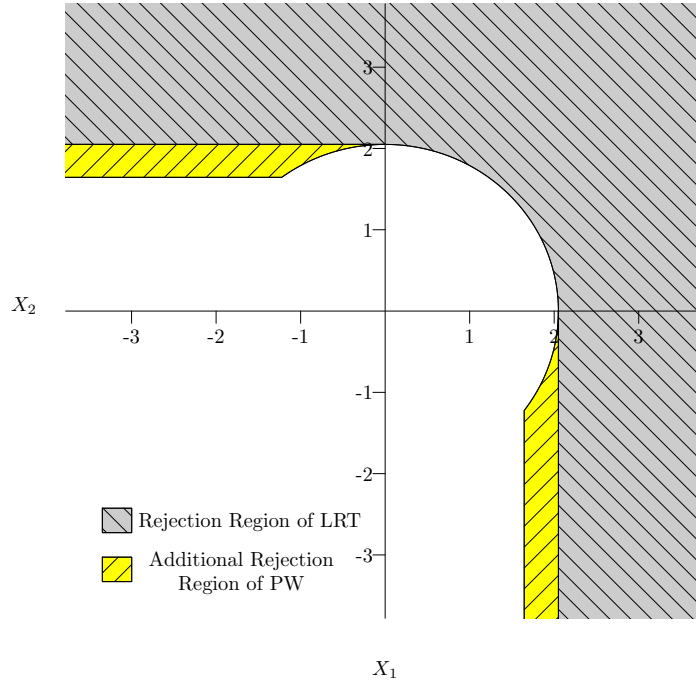


Figure 1: Rejection regions of LRT and PW test.

The data under consideration are assumed to be a random sample from $m + 1$ populations with some clustered structure:

$$\mathbf{y}_{k,j}^T = (y_{k,j,1}, \dots, y_{k,j,d}) : k = 0, 1, \dots, m; j = 1, 2, \dots, n_k.$$

In this setting, k is the identity of the population, d is the cluster size, and n_k is the number of clusters sampled from the k th population.

Let $F_k(\mathbf{y})$ be the cumulative joint distribution (CDF) of $\mathbf{y}_{k,j}$. The nature of the data implies that F_k is exchangeable. The exchangeability implies an identical marginal distribution, which will be denoted $G_k(y)$. The target of the monitoring test is hence $G_k(y)$. We wish to be alerted when $G_k(y)$ is stochastically smaller than $G_0(y)$ in some respect. As

pointed out earlier, we may test if G_k is lower than G_0 in the 5% quantile or the median.

Because the G_k 's are of a similar nature, Chen et al. (2016) suggested that the density ratio model (DRM) (Anderson 1979) is appropriate. Specifically, they assumed that these distributions are related through the following equation:

$$\frac{dG_k(y)}{dG_0(y)} = \exp\{\boldsymbol{\beta}_k^T \mathbf{q}(y)\} \quad (12)$$

for a suitably selected function $\mathbf{q}(y)$ of dimension q with unknown parameter vectors $\boldsymbol{\beta}_k$.

Based on the DRM, Chen et al. (2016) proposed the following composite empirical likelihood (EL):

$$L(G_0, \boldsymbol{\beta}) = \prod_{k,j,l} \{dG_k(y_{k,j,l})\} = \left(\prod_{k,j,l} p_{k,j,l} \right) \exp \left\{ \sum_{k,j,l} \boldsymbol{\beta}_k^T \mathbf{q}(y_{k,j,l}) \right\} \quad (13)$$

where $G_0(y) = \sum_{k,j,l} p_{k,j,l} 1(y_{k,j,l} \leq y)$. The DRM assumption implies

$$\sum_{k,j,l} p_{k,j,l} \exp\{\boldsymbol{\beta}_r^T \mathbf{q}(y_{k,j,l})\} = 0 \quad (14)$$

for $r = 0, 1, \dots, m$.

Some algebra shows that the above composite EL has a dual form:

$$\ell_n(\boldsymbol{\beta}) = - \sum_{k,j,l} \log \left[\sum_{r=0}^m \rho_r \exp\{\boldsymbol{\beta}_r^T \mathbf{q}(y_{k,j,l})\} \right] + \sum_{k,j,l} \boldsymbol{\beta}_k^T \mathbf{q}(y_{k,j,l}). \quad (15)$$

Many of the numerical computations are done via the dual form.

Let the maximum composite EL estimator be $\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})$. Let

$$\hat{G}_r(y) = \sum_{k,j,l} \hat{p}_{k,j,l} \exp\{\hat{\boldsymbol{\beta}}_r^T \mathbf{q}(y_{k,j,l})\} \mathbb{1}(y_{k,j,l} < y)$$

be the fitted CDF, with the obvious notation $\hat{p}_{k,j,l}$. By the invariance property of the maximum likelihood estimation, we estimate the population means and quantiles by

$$\hat{\mu}_r = \sum_{k,j,l} \hat{p}_{k,j,l} \exp\{\hat{\boldsymbol{\beta}}_r^T \mathbf{q}(y_{k,j,l})\} y_{k,j,l}$$

and

$$\hat{\xi}_r = \hat{\xi}_{r,\alpha} = \inf\{y : \hat{G}_r(y) \geq \alpha\}$$

where α denotes the level of the quantile. It has been shown that the parameter estimators are asymptotically normal. For instance, in obvious notation,

$$\sqrt{n}\{(\hat{\xi}_1, \hat{\xi}_2) - (\xi_1, \xi_2)\} \rightarrow N(0, \mathbf{\Sigma}).$$

A cluster-based bootstrap method proposed by Chen et al. (2016) can be used for the consistent estimation of $\mathbf{\Sigma}$.

We are now ready to apply the modified LR test to the one-sided test problem for multiple parameters. Suppose $\boldsymbol{\theta}$ is a vector-valued parameter. Let $\hat{\boldsymbol{\theta}}$ be its MLE and \mathbf{S}^* be its bootstrap variance estimator given in Chen et al. (2016). The monitoring test problem is transformed to the problem of testing for some hypothesis in the form of (1). When

$$\boldsymbol{\theta} = (\xi_{1,0.05} - \xi_{0,0.05}, \xi_{1,0.50} - \xi_{0,0.50})^T,$$

testing for (1) involves monitoring whether G_1 has simultaneously maintained the 5th percentile and the median of the wood strength distribution compared to G_0 . In the presence of multiple populations, the test is more efficient if we also utilize information from G_2 , G_3 , and so on (Chen et al. 2016). Depending on the monitoring target, other forms of $\boldsymbol{\theta}$ can easily be specified.

The null hypothesis of interest is $\boldsymbol{\theta} \geq 0$. To apply the proposed modified LRT, we compute the value of T_n given in (4) with

$$\mathbf{X} = -\hat{\boldsymbol{\theta}}; \quad \mathbf{S} = n\mathbf{S}^*.$$

The reason for the negative sign in $\mathbf{X} = -\hat{\boldsymbol{\theta}}$ is to reconcile the opposite inequalities specified in (1) and (2). We compute the p-value of the test according to (10). Clearly, we could as easily use other tests based on \mathbf{X} and \mathbf{S} .

4 Simulation and example

In this section, we use simulation to discover the pros and cons of three tests: LRT, PW, and the proposed mLR for one-sided hypotheses. We do not include UIT because this method has been shown to be inferior by Perlman & Wu (2003) and Perlman & Wu (2006). As pointed out earlier, the type I errors of the mLR and PW tests likely exceed the desired size for some distributions. It is important to explore how serious the errors become and the features of the corresponding distributions.

We focus on the situation where the dimension of the parameter $p = 2$ with a sample of size $n = 50$ from various multivariate normal distributions.

4.1 Multivariate normal samples

It can easily be seen that the test problem of interest is invariant to the variance of the marginal distributions. When $p = 2$, this implies that we need consider only the covariance matrices in the following form:

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We generated data from 4×5 null models with a range of correlation coefficients:

$$\rho = -0.9, -0.5, 0.0, 0.5, 0.9.$$

From each model, we generated 100,000 samples of size $n = 50$. We set the nominal rejection rate, or size of the test, to 0.05. The values of the population mean $\boldsymbol{\mu}$ and the percentage of times when the null hypothesis is rejected by these four tests are summarized in Table 2.

Null models. Let us first examine the results for $\boldsymbol{\mu} = (0, 0)^T$ at which the null hypothesis is true. The results in Table 2 support the theory that LRT and UIT tightly

control the type I error. However, they achieve this goal by being very conservative at $\rho = 0.5, 0.9$. The PW test improves on LRT and UIT in terms of being less conservative, but at the cost of exceeding the nominal level at $\rho = -0.9$. The type I errors of the proposed mLR over this range of ρ are very close to the nominal level.

When $\boldsymbol{\mu}$ goes from $(0, -0.1)$ to $(0, -0.3)$, the null hypothesis remains true. Since it makes the model move toward the “interior” of H_0 , the type I errors of these tests become lower, as expected.

Alternative models. We also carry out simulation for three sets of alternative distributions. In the first, both marginal means become greater than 0 at the same rate. In the second, just one of the marginal means becomes greater than 0. In the third, two marginal means move in opposite direction. The simulated powers of the three tests are given in the second, third and fourth blocks of Table 2.

Clearly, LRT has lower power than PW and mLR for the alternative distributions. The comparison between PW and mLR is not clear-cut: mLR is uniformly more powerful than PW for the first set of alternative distributions (second block of Table 2). For the second set (third block of Table 2) mLR has higher power than PW when $\rho = -0.9, -0.5$, and 0; comparable power when $\rho = 0.5$; and slightly lower power when $\rho = 0.9$. For the third set (fourth block of Table 2) PW is more powerful.

Based on the simulation results, we recommend using the PW test in applications where the two quality indices may move in opposite directions. If the two indices are likely to move in the same direction, mLR is preferable.

4.2 Application to multiple quality indices in monitoring context

We now study the use of the proposed test for multi-dimensional quality indices in monitoring. We simulate data with a cluster structure, as discussed in Section 3. We compare

Table 2: Type I errors for one-sided tests (%)

	$\rho = -0.9$			$\rho = -0.5$			$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
μ	LRT	PW	mLR	LRT	PW	mLR	LRT	PW	mLR	LRT	PW	mLR	LRT	PW	mLR
(0,0)	4.49	5.46	5.08	3.79	4.79	5.05	3.05	4.03	4.98	2.46	3.43	5.00	1.66	2.64	4.94
(0,-.1)	1.34	4.23	1.56	1.54	3.03	2.12	1.62	2.91	2.69	1.34	2.70	2.87	1.18	3.62	3.65
(0,-.2)	1.23	5.13	1.40	1.11	3.60	1.56	1.20	3.32	2.04	1.20	3.69	2.59	1.15	4.90	3.55
(0,-.3)	1.14	5.12	1.31	1.14	4.48	1.58	1.21	4.28	2.05	1.18	4.50	2.55	1.13	4.89	3.45
(.1,.1)	84.4	84.4	85.7	27.6	27.9	32.4	16.1	17.1	22.3	11.2	12.9	18.8	7.74	10.2	17.3
(.2,.2)	100	100	100	75.8	75.8	80.1	46.8	47.3	56.3	32.4	34.1	45.3	23.6	27.5	40.8
(.3,.3)	100	100	100	97.9	97.9	98.6	80.3	80.4	86.2	62.2	63.1	74.3	49.4	53.4	68.7
(.4,.4)	100	100	100	99.9	99.9	99.9	96.3	96.3	97.9	86.2	86.5	92.5	75.1	77.6	88.3
(0,.1)	33.9	34.1	36.0	12.4	13.4	15.5	8.88	10.6	13.0	6.79	9.64	12.4	5.72	12.6	13.5
(0,.2)	86.1	86.1	87.3	33.4	34.3	38.8	23.1	26.3	30.4	19.4	27.5	29.6	18.4	38.9	33.9
(0,.3)	99.5	99.5	99.6	62.7	63.1	68.0	46.7	50.8	55.9	42.3	56.0	55.6	41.9	67.2	61.3
(0,.4)	100	100	100	86.4	86.6	89.3	72.0	75.3	79.2	68.8	82.0	79.7	68.2	87.3	83.6
(-.1,.1)	8.19	10.3	9.04	7.07	9.89	9.14	6.26	10.1	9.44	5.83	11.7	10.6	5.53	16.7	13.4
(-.2,.2)	21.1	25.9	22.7	19.6	27.8	23.6	18.7	31.4	25.2	18.4	37.5	28.4	18.6	40.5	34.4
(-.3,.3)	44.0	51.2	46.1	42.1	55.9	47.5	41.8	62.8	50.9	41.7	66.8	54.8	41.6	67.1	61.3
(-.4,.4)	69.3	76.2	71.2	68.5	81.8	73.2	68.3	86.2	75.9	68.3	87.4	79.3	68.5	87.4	83.7

the LRT and PW test and we again omit UIT.

We consider the situation where clustered random samples from $m + 1 = 4$ populations are available and the cluster size $d = 5$. We use $B = 999$ bootstrap repetitions for the variance estimation. To paint a more complete picture, we simulated data from two clustered population sets: one is multivariate normal and the other is multivariate gamma. The reliability literature indicates that these are sensible models for data from quality indices. We emphasize that the data analysis does not assume knowledge of the data-generating distributions.

Multivariate clustered normal populations We first perform simulation by generating individual response values from the following random effect model:

$$y_{k,j,l} = \mu_k + \gamma_{kj} + \epsilon_{kjl}.$$

In the wood product application, $y_{k,j,l}$ is the mechanical strength of a piece of wood from the k th population, j th cluster, and l th unit. We generate γ_{kj} from $N(0, \sigma_{\gamma,k}^2)$. Since γ_{kj} is shared by all the units in cluster j in the k th population, it induces within-cluster positive correlation. We generate ϵ_{kjl} from $N(0, \sigma_e^2)$, which reflects the noise in the mechanical strength. The marginal distributions G_k are all normal, but this fact will not be used in the hypothesis test. Instead, we use DRM with $\mathbf{q}(y) = (1, y, y^2)^T$.

The problem of interest in the targeted application is whether or not the 5th percentile and the median of the mechanical strength of year $k > 0$ are maintained compared to some base year $k = 0$. Let $\xi_{k,\alpha}$ be the α th percentile of G_k . Let

$$\boldsymbol{\theta}_k = (\xi_{k,0.05} - \xi_{0,0.05}; \xi_{k,0.50} - \xi_{0,0.50}).$$

For the purposes of illustration, we test, for each $k = 1, 2, 3$ not simultaneously,

$$H_0 : \boldsymbol{\theta}_k \geq 0 \text{ against the alternative } H_a : \boldsymbol{\theta}_k \not\geq 0.$$

Clearly, the proposed test can be used for any other suitable quality indices. The same is true for the LRT and the PW test.

The simulation was conducted with three sets of parameters:

	(μ_0, \dots, μ_3)	$(\sigma_{\gamma,0}, \dots, \sigma_{\gamma,3})$	σ_e	Feature
I	(15.5, 15.5, 14.7, 14.0)	(1.2, 1.2, 1.0, 1.0)	2.0	$\xi_{.05}, \xi_{.50}$ reduced
II	(15.5, 15.2, 15.0, 14.7)	(2.0, 1.794, 1.653, 1.436)	1.0	$\xi_{.05}$ reduced
III	(15.5, 15.5, 15.5, 15.5)	(1.0, 1.2, 1.4, 1.6)	1.0	$\xi_{.50}$ reduced

The numbers of clusters are chosen to be $(n_0, n_1, n_2, n_3) = (25, 30, 40, 40)$. The quantile and median values are given by

	$(\xi_{0,0.5}, \dots, \xi_{3,0.5})$	$(\xi_{0,0.05}, \dots, \xi_{3,0.05})$
I	(15.50, 15.50, 14.70, 14.00)	(11.66, 11.66, 11.02, 10.32)
II	(15.50, 15.20, 15.00, 14.70)	(11.82, 11.82, 11.82, 11.82)
III	(15.50, 15.50, 15.50, 15.50)	(13.17, 12.93, 12.67, 12.40)

In the first setting, the first two populations are identical and the other two populations have a lower 5th percentile and median. This arrangement allows us to investigate the type I error by testing $\theta_1 \geq 0$ and the power for $\theta_2 \geq 0$ and $\theta_3 \geq 0$. In the second setting, the four populations have the same median, but the 5th percentile reduces from the first to the last population. In the third setting, the four populations have the same 5th percentile, but the median reduces from the first to the last population.

We set the number of repetitions to 10,000. The simulated rejection rates for the three hypotheses are summarized in Table 3.

Recall that in Setting I, the null hypothesis $\theta_1 \geq 0$ is true. The simulation results clearly show that the faithful LRT has a much lower type I error than the nominal size of 5%. This is not bad in itself. The problem is that the lower type I error is at the cost

Table 3: Simulated rejection rates for normal data (%)

	Setting I			Setting II			Setting III		
H_0	LRT	PW	mLR	LRT	PW	mLR	LRT	PW	mLR
$\boldsymbol{\theta}_1 \geq 0$	2.93	3.86	5.91	4.50	5.20	8.20	6.29	9.34	12.08
$\boldsymbol{\theta}_2 \geq 0$	47.35	52.55	62.01	7.20	10.30	13.30	16.45	25.55	25.87
$\boldsymbol{\theta}_3 \geq 0$	95.81	96.83	98.20	14.30	22.20	24.70	29.18	44.44	42.93

of a much lower power for rejecting $\boldsymbol{\theta}_2 \geq 0$ and $\boldsymbol{\theta}_3 \geq 0$ compared to the other methods. Comparing PW and mLR shows that PW is also too conservative and therefore has low power. The mLR has higher power but also higher type I error.

The null hypotheses for Settings II and III are false, and so power is measured by the rejection of the hypothesis. The simulation results in Table 3 generally favor mLR. Overall, we conclude that the proposed mLR works well.

Multivariate clustered gamma populations We now perform simulation by generating individual response values from multivariate clustered gamma populations.

One way to create multivariate clustered gamma observations is as follows. Let U_1, \dots, U_d be d iid random variables following beta distributions with shape parameters a and b . Further, let W be a gamma-distributed random variable with shape parameter $a + b$ and rate parameter β . Then

$$\mathbf{Y} = W(U_1, \dots, U_d)^T$$

is multivariate gamma $MG(a, b, \beta)$ with correlation $\text{cor}(Y_i, Y_j) = a/(a + b)$ for all $1 \leq i < j \leq d$. The marginal distribution of $\mathbf{Y}_1 = U_1 W$ is gamma with shape parameter a and rate parameter β . When $b = \infty$, Y_1, \dots, Y_d become independent; see Nadarajah & Gupta (2006).

The simulation was conducted with three sets of parameters:

	(a_0, \dots, a_3)	$(\beta_0, \dots, \beta_3)$	b	Feature
I	(8.0, 8.0, 7.0, 6.0)	(1.00, 1.00, 1.05, 1.10)	14	$\xi_{.05}, \xi_{.50}$ reduced
II	(8.0, 8.5, 9.0, 10)	(1.00, 1.09, 1.18, 1.36)	14	$\xi_{.05}$ reduced
III	(8.0, 7.0, 6.0, 5.0)	(1, 0.87, 0.74, 0.61)	14	$\xi_{.50}$ reduced

The quantile and median values are given by

	$(\xi_{0,0.5}, \dots, \xi_{3,0.5})$	$(\xi_{0,0.05}, \dots, \xi_{3,0.05})$
I	(7.67, 7.67, 6.35, 5.15)	(3.98, 3.98, 3.13, 2.38)
II	(7.67, 7.49, 7.35, 7.11)	(3.98, 3.98, 3.98, 3.98)
III	(7.67, 7.67, 7.67, 7.67)	(3.98, 3.78, 3.53, 3.23)

We test the same hypotheses as for the multivariate clustered normal populations. The results are given in Table 4.

Table 4: Simulated rejection rates for gamma data (%)

	Setting I			Setting II			Setting III		
H_0	LRT	PW	mLR	LRT	PW	mLR	LRT	PW	mLR
$\theta_1 \geq 0$	2.79	3.76	5.69	76.07	77.48	86.23	99.99	99.99	100.0
$\theta_2 \geq 0$	4.17	5.43	7.96	6.25	8.96	12.30	13.80	21.82	23.51
$\theta_3 \geq 0$	6.01	8.72	11.27	14.21	21.17	22.84	32.87	47.61	45.44

Our observations are similar to those for the multivariate clustered normal populations. Both LRT and PW are too conservative: the type I error is much lower than 5% in Setting I, for the null hypothesis $\theta_1 \geq 0$. The PW test is also too conservative and therefore has low power. The mLR has higher power but also higher type I error. The overall impression is that the proposed mLR works well.

4.3 Data analysis

We now apply our method to a real forestry data set. It contains 398 modulus of rupture (MOR) measurements from In-Grade samples and 408 MOR measurements from monitoring samples obtained in 2011/2012. Both Chen et al. (2016) and Verrill et al. (2015) found that the 5th quantile is markedly reduced in the monitoring sample with high statistical significance. We certainly expect that any one-sided hypothesis tests for the 5th quantile

Table 5: Sample quantiles of forestry data

	5%	50%
In-Grade	2.64	5.28
2011/2012	1.87	3.71

and the median of MOR will produce a statistically significant outcome. In this analysis, we used the basis function $\mathbf{q}(y) = (1, y, y^2, \log y)$ suggested by Chen et al. (2016). The estimated differences in the 5th quantile and the median are $(\hat{\theta}_{0,1;0.05}, \hat{\theta}_{0,1;0.5}) = (-0.69, -1.53)$. By the bootstrap method recommended by Chen et al. (2016), the asymptotic covariance matrix of this estimator is estimated as

$$\mathbf{S}_n = \begin{pmatrix} 0.01282 & 0.01586 \\ 0.01586 & 0.04022 \end{pmatrix}.$$

We now use $\mathbf{X} = (0.69, 1.53)^T$ and $\mathbf{S} = n\mathbf{S}_n$ to compute T_n defined in (4). We find $T_n = 59.3$ and $\hat{p} = 2.30 \times 10^{-14}$ by (10). Hence, the null hypothesis is rejected with strong statistical evidence.

Note that the estimated correlation coefficient is $\hat{\rho} = 0.70$ in this example. This is the value used to compute \hat{p} . When the LRT is applied to this problem, we compute the p-value as if $\rho = -1$, giving 7.15×10^{-14} . The result remains sufficiently significant, but

there is a large drop in the level of significance. The p-value of the PW test is the same in this case.

The two populations in this example are so different that the quality deterioration is detected by any reasonable methods. To demonstrate more subtle differences between methods, we artificially inflate every data point of the 2011/2012 sample by a factor of 1.35. The two samples now have much closer sample-quality indices: the estimated differences in the 5th quantile and the median are $(\hat{\theta}_{0,1;0.05}, \hat{\theta}_{0,1;0.5}) = (-0.166, -0.009)$. The estimated asymptotic covariance matrix of this estimator is

$$\mathbf{S}_n = \begin{pmatrix} 0.0081 & 0.0156 \\ 0.0156 & 0.0545 \end{pmatrix}.$$

We now find $T_n = 3.41$, and the p-values based on LRT, PW, and mLR are 0.123, 0.032, and 0.053. Because the change in the median is so small, the PW test arrives at its p-value primarily because of the large $|\hat{\theta}_{0,1;0.05}|$. In comparison, mLR takes a more balanced view of the two indices, and the differences in the median and 5% quantile between the two populations are judged not significant at the 5% level. The LRT is too conservative, as our simulations predicted.

5 Conclusions

One-sided multi-parameter hypothesis tests arise in many applications, and there are many effective test methods under normal models with a solid theoretical basis. We are particularly interested in testing whether two quality indices are reduced over time. The existing methods have room for further improvement, particularly in the context of our application. We propose a new test for this context. In particular, we have developed a strategy for applying the method to general one-sided multi-parameter hypotheses.

References

- Anderson, J. (1979), ‘Multivariate logistic compounds’, *Biometrika* **66**(1), 17–26.
- ASTM (1991), ‘Standard practice for establishing allowable properties for visually-graded dimension lumber from in-grade tests of full-size specimens’, *American Society for Testing and Materials, West Conshohocken, PA*, <http://www.astm.org> .
- Berger, R. L. (1982), ‘Multiparameter hypothesis testing and acceptance sampling’, *Technometrics* **24**(4), 295–300.
- Chen, J., Li, P., Liu, Y. & Zidek, J. V. (2016), ‘Monitoring test under nonparametric random effects model’, *arXiv preprint arXiv:1610.05809* .
- Kendall, M. G. (1941), ‘Proof of relations connected with the tetrachoric series and its generalization’, *Biometrika* **32**(2), 196–198.
- Lachin, J. M. (2014), ‘Applications of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales’, *PloS one* **9**(10), e108784.
- Nadarajah, S. & Gupta, A. K. (2006), ‘Some bivariate gamma distributions’, *Applied Mathematics Letters* **19**(8), 767–774.
- Nüesch, P. E. (1966), ‘On the problem of testing location in multivariate populations for restricted alternatives’, *The Annals of Mathematical Statistics* **37**(1), 113–119.
- O’Brien, P. C. (1984), ‘Procedures for comparing samples with multiple endpoints’, *Biometrics* **40**(4), 1079–1087.
- Perlman, M. D. (1969), ‘One-sided testing problems in multivariate analysis’, *The Annals of Mathematical Statistics* **40**(2), 549–567.

- Perlman, M. D. & Wu, L. (2003), ‘On the validity of the likelihood ratio and maximum likelihood methods’, *Journal of Statistical Planning and Inference* **117**(1), 59–81.
- Perlman, M. D. & Wu, L. (2006), ‘Some improved tests for multivariate one-sided hypotheses’, *Metrika* **64**(1), 23–39.
- Robertson, T. & Robertson, T. (1988), Order restricted statistical inference, Technical report.
- Roy, S. N. (1953), ‘On a heuristic method of test construction and its use in multivariate analysis’, *The Annals of Mathematical Statistics* **24**(2), 220–238.
- Tamhane, A. C. & Logan, B. R. (2004), ‘A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials’, *Biometrika* **91**(3), 715–727.
- Tang, D.-I., Gnecco, C. & Geller, N. L. (1989), ‘An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials’, *Biometrika* **76**(3), 577–583.
- Verrill, S., Kretschmann, D. E. & Evans, J. W. (2015), Simulations of strength property monitoring tests. Unpublished manuscript. Forest Products Laboratory, Madison, Wisconsin. Available at <http://www1.fpl.fs.fed.us/monit.pdf>.
- Wassmer, G., Reitmeir, P., Kieser, M. & Lehmacher, W. (1999), ‘Procedures for testing multiple endpoints in clinical trials: an overview’, *Journal of statistical planning and inference* **82**(1), 69–81.